WEIR-P: An Information Extraction Pipeline for the Wastewater Domain

Nanée Chahinian, Thierry Bonnabaud La Bruyère, Francesca Frontini, Carole Delenne, Marin Julien, Rachel Panckhurst, Mathieu Roche, Lucile Sautot, Laurent Deruelle, Maguelonne Teissiere

WEIR-P: An Information Extraction Pipeline for the Wastewater Domain*

Nanée Chahinian[1], Thierry Bonnabaud La Bruyère[1], Francesca Frontini[2,3], Carole Delenne[1,4], Marin Julien[1], Rachel Panckhurst[5], Mathieu Roche[6], Lucile Sautot[6], Laurent Deruelle[7], and Maguelonne Teissiere[6]

[1] HSM, Univ. Montpellier, CNRS, IRD, France
nanee.chahinian@ird.fr
[2] Instituto di Linguistica Computazionale "A. Zampolli" - CNR, Pisa, Italy
[3] CLARIN ERIC
[4] Inria Lemon, CRISAM - Inria Sophia Antipolis – Méditerranée, France
[5] Dipralang, UPVM, Montpellier, France
[6] TETIS, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[7] Berger Levrault, Perols, France

Abstract. We present the MeDO project, aimed at developing resources for text mining and information extraction in the wastewater domain. We developed a specific Natural Language Processing (NLP) pipeline named WEIR-P (WastewatEr InfoRmation extraction Platform) which identifies the entities and relations to be extracted from texts, pertaining to network information, wastewater treatment, accidents and works, organizations, spatio-temporal information, measures and water quality. We present and evaluate the first version of the NLP system which was developed to automate the extraction of the aforementioned annotation from texts and its integration with existing domain knowledge. The preliminary results obtained on the Montpellier corpus are encouraging and show how a mix of supervised and rule-based techniques can be used to extract useful information and reconstruct the various phases of the extension of a given wastewater network. While the NLP and Information Extraction (IE) methods used are state of the art, the novelty of our work lies in their adaptation to the domain, and in particular in the wastewater management conceptual model, which defines the relations between entities. French resources are less developed in the NLP community than English ones. The datasets obtained in this project are another original aspect of this work.

Keywords: Wastewater · Text mining · Information extraction · NLP · NER · Domain adapted systems.

1 Introduction

Water networks are part of the urban infrastructure and with increasing urbanization, city managers have had to constantly extend water access and sanitation

* Supported by the Occitanie-Pyrénées-Méditerrannée Region under grant 2017-006570/DF-000014.

services to new peripheral areas or to new incomers [3]. Originally these networks were installed, operated, and repaired by their owners [24]. However, as concessions were increasingly granted to private companies and new tenders requested regularly by public authorities, archives were sometimes misplaced and event logs were lost. Thus, part of the networks' operational history was thought to be permanently erased. However, the advent of Web big data and text-mining techniques may offer the possibility of recovering some of this knowledge by crawling secondary information sources, i.e. documents available on the Web. Thus, insight might be gained on the wastewater collection scheme, the treatment processes, the network's geometry and events (accidents, shortages) which may have affected these facilities and amenities. This is the primary aim of the "Megadata, Linked Data and Data Mining for Wastewater Networks" (MeDo) project, funded by the Occitanie Region, in France, and carried out in collaboration between hydrologists, computational linguists and computer scientists.

Text mining is used in the field of water sciences but it is often implemented for perception analysis [18, 1], indicator categorization [22] or ecosystem management [10, 14]. To the best of our knowledge this work is the first attempt at using Natural Language Processing (NLP) and Text-Mining techniques for wastewater network management in cities.

The creation of domain adapted systems for Information Extraction (IE) has been an important area of research in NLP. The outcomes of many projects have led to creating systems capable of identifying relevant information from scientific literature as well as technical and non technical documents. A pioneering domain was that of biology; see [2] for an overview of initial projects and [11] for a more recent contribution. More specifically, a fertile field of research has emerged at the intersection between NLP and Geographic Information Retrieval (GIR). Cf. [30] and [19] for a domain overview and a list of relevant projects; [16] for Digital Humanities initiatives and [13] for the Matriciel project in French.

The NLP pipelines used to mine these specialised corpora are generally composed of a mix of supervised and unsupervised or rule-based NLP algorithms; crucially the collaboration with domain experts is essential since the state of the art tools need to be re-trained and tested using corpora with specific annotation and often require structured domain knowledge as input. Generally such systems are trained and tested on English; however in many cases multilingual pipelines, and therefore resources, have been created.

The use of general semantic resources e.g. EuroWordNet[8] or dedicated thesaurii Agrovoc[9] for the French language is not adapted to address the wastewater domain.Agrovoc is a terminology, where concepts are described and translated into various languages. In this sense it is used to standardize and control the use of domain vocabulary. WordNets are lexical resources, that can be used for NLP; EuroWordnet has been extended for specific domains. In both cases, to produce or adapt terminologies or computational lexicons for a new domain one would have to apply NLP techniques to extract and filter terms and lexemes from

[8] https://archive.illc.uva.nl//EuroWordNet/
[9] http://www.fao.org/agrovoc/about

texts, and use a pre-processing pipeline along the lines of what we propose. The originality of our approach is based on the combination of well-known tools (e.g. spaCy, Heideltime) and the integration of expertise for highlighting semantic information related to the wastewater domain along with the selection of relevant documents using machine learning. Our global pipeline called WEIR-P combines Information Retrieval (IR) and Information Extraction (IE) techniques adapted for the French language and the wastewater domain.

Indeed, not only are there no domain specific terminological and lexical resources ready to use, but crucially also in terms of modelling, important work had to be done to identify which elements had to be annotated in the text, in order to extract the relevant information. In particular, two important contributions of this work are the implementation of NER methods for extracting original entities (Network element, Treatment, Network type, Accident, etc.) with new textual data manually labeled and the construction of relations based on spatial, temporal and thematic entities for discovering new knowledge is an original research direction. Finally French resources (i.e. corpora, specialized terminology, etc.) are less developed in the NLP community with many English datasets available. The datasets obtained in this project (i.e. labeled corpus, terminology dedicated to wastewater domain) are another original aspect of this work.

The **research methodology** we followed is quite similar to the DSRP model presented by [23]. The **problem identification and motivation is linked to previous research projects [5]. The Objectives of the solution are to develop a user-friendly tool that would help non-IT researchers and wastewater network managers transform unstructured text data from the Web into structured data and merge them with their business databases. The Design and Development step relied heavily on NLP, IR and IE literature. The new French data standard for drinking water and sewerage networks issued halfway into the project [12] was used as a reference for the business data model. Two case studies were designed in two cities (Montpellier-France and Abidjan-Ivory Coast). Demonstrations were carried out during the meetings of the project's steering committee. The Evaluation is two fold. It was partly carried out during the bi-annual meetings of the steering committee. A restitution workshop is also planned with members of the Aqua-Valley French Competitiveness Water Cluster. As for the Communication, in addition to scholarly publications such as this one, a wastewater awareness workshop will be carried out in a primary school in March 2021.**

In this paper we shall describe the general architecture of the system, as well as the linguistic resources used for domain adaptation and the evaluation of the various steps of the NLP pipeline. The structure of the paper is as follows: description of the data model (§2), of the system's architecture (§3) and its evaluation (§4); finally the description of the resources produced (§5) and some conclusions (§6).

2 The Data model and the corpus

The MeDo project aims to use textual data – available on the Web or produced by institutions – for learning about the geometry and history of wastewater networks, by combining different data-mining techniques and multiplying analysed sources.

In order to achieve this, a system for document processing has been designed, which when in production will allow a hydrologist or a wastewater network manager to retrieve relevant documents for a given network, process them to extract potentially new information, assess this information by using interactive visualization and add it to a pre-existing knowledge base.

Although textual sources may be directly imported into the system, the default pipeline begins with a Web-scraping phase retrieving documents from the Internet. The system has been built within the context of a French regional project, and is currently developed and tested for French documents. However, its architecture allows for easy adaptation to other languages, including English.

Relevant information may come from technical documents, such as reports by municipalities or private companies, but also from the general public, for instance from newspaper articles and social media posts containing descriptions of accidents in the network, or announcements of works. The goal of the system is to allow for the identification and retrieval of relevant documents from the Web, their linguistic processing and the extraction of domain information.

The case study presented hereafter is on the city of Montpellier in France. Our corpus is composed of 1,557 HTML and PDF documents. The documents were collected in July 2018 using a set of Google queries with a combination of keywords. Given the absence of guidelines for the annotation and extraction of information for wastewater management we proceeded to create our own annotation model with the help of hydrologists [7]. As is generally the case, the MeDo domain specific annotation scheme has been created extending the commonly used Named Entities (NEs) tagsets, which are defined in the guidelines of well known annotation campaigns such as MUC-5 and following [9], with domain specific entities, or tags. Our current model contains a set of entities:

- three of them are specific to the network: **Network type**, **Network element** and **Network characteristics**. Words such as “réseau pluvial” (stormwater network), “collecteur principal” (sewer main) or “gravitaire” (gravity fed) fall into these categories;
- one is related to the **Wastewater treatment** and may be used either for the plant and its components or the treatment process. This category includes for instance words like “station d’épuration” (wastewater treatment plant), “digestion anaérobie” (anaerobic digestion);
- two entities are used for the type of event reported in the document: **Accident**, **Works** i.e. “pollution”, “inondation” (inundation/flooding), “raccordement au réseau” (connection to the network);
- one label describes the public or private **Organizations** such as “commune” (City council), “Entreprise” (company);

- dates and locations are marked using two labels: **Temporal**, **Spatial**;
- all quantitative data of relevance are marked as **Measurements**;
- a category **Indicator** is used to annotate any information which specifies or adds to the information provided by other categories;
- finally, two more "generic" entities are used for all words related to **Water quality** i.e. "eau brute" (raw water supply), nitrates and the **non-technical or legal aspects of wastewater management** i.e. "délégation de service public" (Public service delegation), "directive cadre sur l'eau" (Water framework Directive).

The manual annotation of these entities is complex and requires expert knowledge to be carried out correctly. A test on **inter-annotator agreement** was carried out with two experts, who had in-depth knowledge of the annotations guidelines, and non experts - students who received one-hour training before annotating. The former were asked to annotate all categories, the latter were limited to the domain specific ones. Results (see Table 1) are encouraging, given the number and technical nature of categories. It is known that Fleiss' Kappa is not appropriate for multi-entity textual annotation [27]; we therefore evaluate agreement in terms of F-measure values, reporting the results for each typology of annotators. Non experts performed slightly more poorly, which is understandable; typically, they tended to miss more annotations; however, when they did annotate a portion of text, they tended to choose the right category, which seems to indicate that the chosen annotation model is sound. For two domain categories (Water Quality and Accident), which are quite rare in the corpus, agreement remains low. The annotation guide was modified following this experiment to further clarify the annotation guidelines. The documents annotated by the students were not used to train the NER modules.

Given these results, the **Wastewater Domain Gold Standard Corpus** was carefully checked and verified by hydrologists. It consists of 23 manually annotated documents: 3 calls for tenders, 6 announcements from the city's website and 12 newspaper articles, 1 technical document and 1 tweet. They amount to 1,387 sentences and 4,505 entities. The annotation was carried out using the BRAT system[10]. Figure 1 shows an example of an annotated sentence. An extended Gold Standard containing 80 documents with 649,593 words and 29,585 entities was later complied for further tests. It is available on DataSuds dataverse [6].

Once the entities have been identified, it is necessary to define the possible **relations** which may connect them. These are defined in the **Wastewater Management Conceptual Model** (see Figure 2) which was developed in collaboration with hydrologists.

Based on the possible abstract relations between the entity classes in the model, entity instances found in texts are thus linked with relations:an event of type *Works* is *spatially* and *temporally* localised and specified by the *type of intervention* (implementing a new wastewater network).

[10] https://brat.nlplab.org/

Table 1. Agreement per category, calculated using F-Measure.

Category	Expert	Non Expert
Network_element	0.875	0.688
Works	0.875	0.719
Treatment	0.750	0.703
Network_type	0.750	0.719
Indicator	0.875	0.484
Network_management	0.750	0.609
Network_characteristics	0.875	0.656
Water_quality	0.375	0.422
Accident	0.250	0.328
Measure	0.875	0.680
Spatial	0.875	-
Temporal	0.875	-
Organization	0.875	-
Overall	0.875	0.719

Fig. 1. A portion of the corpus manually annotated for Named Entities.
English: On March 13, the road to Murviel will be closed to allow works to create a wastewater network

3 The WEIR-P system: Architecture

To automatically extract items of information such as the ones previously described, an information extraction system has been created. The NLP pipeline is similar to those generally used for this type of task (see references in §1), but had to be adapted for our specific needs. In particular it was necessary to create a system which is able to retrieve relevant sources within a specific geographical area, to manage different types of documents, to detect and correlate specific information in texts and finally to relate such information with existing knowledge on a given wastewater system. For this reason a pluri-disciplinary approach was required.

The global architecture of the WEIR-P system is depicted in Figure 3. The main steps of the proposed methodology are:

1. Collection of documents;
2. Named-Entity recognition;
3. Semantic relation extraction;
4. Mapping and data visualization.

In this article, we mainly focus on the text mining part of the MeDo project and on the domain adaptation of existing tools for the wastewater domain. It

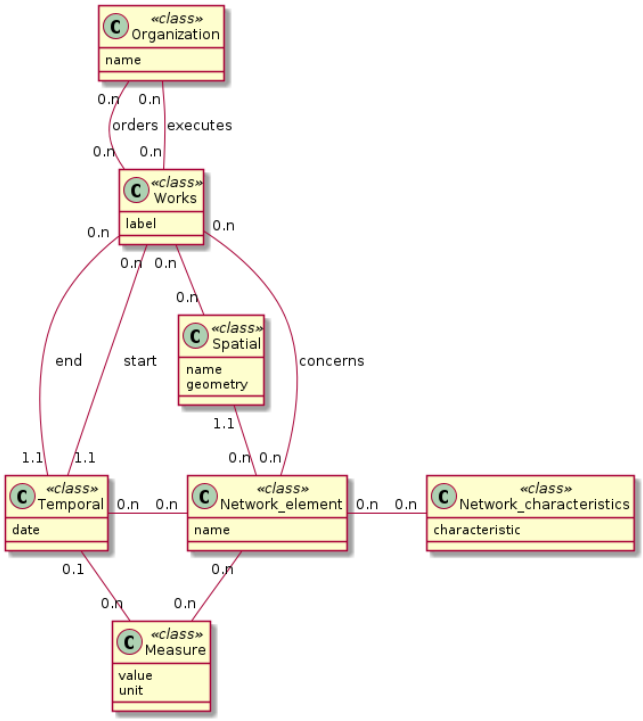

Fig. 2. The wastewater conceptual model.

concerns the 1st, 2nd and 3rd steps of the global architecture of the WEIR-P tool (see Figure 3, Parts A and B).We also briefly present the visualization carried out at document level. The visualization of the network elements on a wastewater map after data fusion will not be presented in this paper.

3.1 Step One - document pre-processing and classification

This step is composed of different sub-steps. The document retrieval algorithm requires the user to specify a geographical area of interest (typically a city or a municipality) as input. The system scrapes documents from the Web using a set of Google queries which specify a combination of two domain-related keywords and a place name.

In the corpus creation phase the texts retrieved from the internet are transformed to plain text using various out-of-the-box Python libraries (such as *BeautifulSoup* or *pdf2txt*); some text cleaning is also necessary, in order to restore broken lines and remove boilerplate text using regular expressions. Finally, all the relevant metadata are recorded and the text collection is entered in the database. Once the corpus has been successfully created, a classification for relevance is carried out using Machine Learning techniques which are detailed in section §4.1.

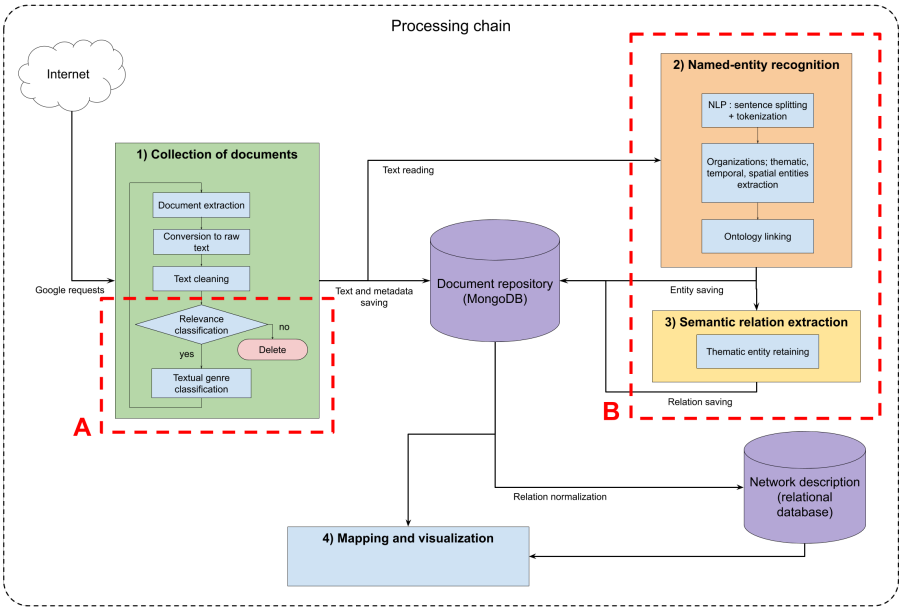

Fig. 3. The WEIR-P Architecture

The pre-processing steps are similar to those followed for text classification using more complex learning methods [29]. This is a necessary step since the crawling algorithm, which is carried out using a coarse filtering system, will retrieve many irrelevant documents which need to be excluded by using a finer-grain classification. For instance query for documents containing words such as “Montpellier”, “eau” (water), “reseau” (network), may sometimes retrieve commercial sites for home plumbing repairs.

3.2 Step Two - Named-Entity Recognition and Classification (NERC)

The goal of the second step is the annotation of Named Entities (NE) in the texts corresponding to the 13 entities or tags defined in §2. The texts are first pre-processed by a basic NLP module which performs sentence splitting and tokenisation and PoS tagging, using the TreeTagger module for French [26]. Subsequently, an **ensemble NERC module** is used, which combines the output of three systems: spaCy [15], CoreNLP [20] and Heideltime [28], which is specifically used for temporal entities.

The system was benchmarked as follows:

1. first spaCy and CoreNLP were trained on a sub-set of the gold standard corpus in order to obtain domain adapted models; to improve the results

for locations, gazetteers of place names (geographical features, locations, addresses) were extracted and used as additional training features from existing geographical bases.
2. then both spaCy and CoreNLP were tested to assess their performance on each class (except temporal entities).
3. Heideltime was tested without training on temporal entities without domain adaptation, and proved to be the best system for this type of entities.

Based on this, the final NERC algorithm runs as follows: first each text is separately annotated with all three systems and the results are compared. For all entities except temporal, in case of conflict between spaCy and CoreNLP, the best performing system for that category is given the priority. For temporal entities the Heideltime annotation is applied, if needed by overriding previous conflicting annotation by the other systems. Results of the evaluation for NERC, as well as the details of the benchmarking for each category are provided in section §4.2.

The NERC module is followed by an **Entity Linking module**, which is used to connect spatial entities such as addresses and locations to existing geographical knowledge bases, in order to produce a cartographic representation of the extracted information. The Geonames[11], BAN[12] and Nominatem[13] geographical databases and corresponding APIs are currently used.

The Entity Linking algorithm is quite basic, since filtering on city or location greatly reduces ambiguity for place names, and will not be further discussed nor evaluated in this paper.

3.3 Step Three - Semantic relation extraction

The objective of the third step is to connect the spatial, temporal and thematic entities discovered in Step Two.

The relation extraction is applied to ensure that the spatial, temporal and network information is accurately linked to the type of event (Works, Accidents), and that network elements and characteristics are correctly linked to each other and the network type, so that they can be used to enrich the knowledge base.

The relation extraction is performed using basic rules. Generally, for this step, we use a more specific textual context (i.e. sentences or paragraphs) than the classification task related to the first step of the MeDo project. To resolve ambiguities in more complex cases we resort to semantics and use syntactic dependencies extracted using a spaCy out-of-the-box dependency parser for French. Preliminary evaluation results are provided in §4.3.

3.4 Step Four - Mapping and Visualization

Step four aims to offer the possibility of visualizing the results of the previous treatments. It consists of two sub-steps and only the first one which deals with

[11] https://www.geonames.org
[12] https://adresse.data.gouv.fr/
[13] https://nominatim.org/

visualization at document level will be presented in this paper. A graphical user interface was developed to enable users to run the annotation pipeline (Steps 1-2-3) on a selected location (e.g. a city), to inspect the extracted entities and relations, and to decide whether to inject the new knowledge in the system. The interface has been developed using FLASK[14] for the Web framework and celery[15] to offset the calculations related to NERC and relation extraction. The d3js library[16] is used to represent the network as a graph where the nodes are coloured according to the NE category (e.g. Temporal, Spatial, Works) and the edges according to the type of relation (e.g. "Has_Temporal", "Has_Spatial"). The Leaflet library[17] is used to map the extracted spatial entities.

The GUI is composed of a standard sign-in module, an administration panel, two monitoring menus to view the progression of tasks and the corresponding notifications and six menus that are more specifically related to the annotation platform. The Corpus menu allows registered users to compile corpora either directly from the , or by uploading a zipped file containing pdf documents. Users may also upload zipped files of previously annotated texts. The "Processing" menu allows users to run the classifications described in section §3.1 and the NE and relation extractions described in section §3.2 either simultaneously or sequentially.

Once these steps are completed, the user has the possibility of visualising the results for each individual document in the "Results" menu (Figure 4). The text and extracted NEs are displayed on the upper part of the screen using the set colour scheme presented in the legend. The semantic relations are displayed as a graph in the lower part of the screen. The export button allows users to create and download a zipped file containing: the document in original and text formats; a pdf of the annotated document and a text file containing the annotations in Brat format (.ann); the graph of relations in svg format; the metadata as a JSON file.

The statistics related to the entire corpus (e.g. number and types of documents, number of words, number and types of entities and relations, number of documents per website, etc.) are visualized in the "Corpus Statistics" menu and exported in pdf format and a map of the extracted spatial entities is displayed in the "Spatial rendering" menu. Finally, the export menu allows users to download a zipped file containing all the documents composing the corpus, the corresponding annotation files in Brat format and a JSON file of the metadata.

4 Evaluation

We describe the evaluation set-up and results of the domain adapted NERC and relation extraction modules.

[14] https://palletsprojects.com/p/flask/
[15] http://www.celeryproject.org/
[16] https ://d3js.org/
[17] https://leafletjs.com/

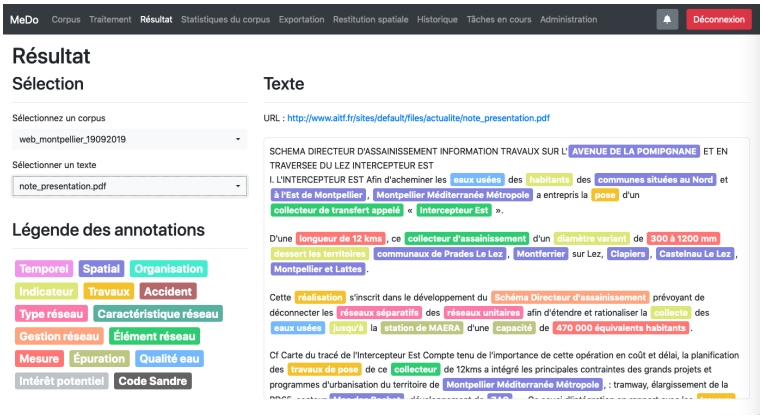

Fig. 4. Screenshot of the results menu.
English: WASTEWATER MASTER PLAN INFORMATION REGARDING WORKS ON AVENUE DE LA POMPIGNANE AND ACROSS THE LEZ RIVER. EASTERN INTERCEPTOR.
1. THE EASTERN INTERCEPTOR. In order to convey the wastewater of the inhabitants of the municipalities located north and east of Montpellier, Montpellier Méditerranée Métropole undertook the installation of a sewer main called "Eastern Interceptor".
With a length of 12 km, this sewer main with diameters varying between 300 to 1200 mm serves the municipal areas of Prades-le-Lez, Montferrier-sur-Lez, Clapiers, Castelnau-le-Lez, Montpellier and Lattes.
This project is part of the Wastewater Master Plan, which provides for the disconnection of the separate networks from the combined sewers in order to extend and rationalise wastewater collection up to the MAERA plant with a capacity of 470,000 equivalent inhabitants.
Cf Map of the Eastern Interceptor. Given the importance of this operation in terms of cost and time, the planning of the installation works for this 12 km sewer main took into account the main constraints of the major urbanization projects and programmes in the Montpellier Méditerranée Métropole area: tramway, widening of

4.1 Text classification systems

In order to learn the classification model, a training corpus was produced using a subset of the Montpellier corpus. Relevant documents should contain information either about network configuration (location, flow type, design rules, material) or incidents (flooding, pollution, water intrusion).

The classifier has been trained using a multinomial Naive Bayes classification method, with a scikit-learn Python library. For each classification the features are bags-of-words, weighted by TF-IDF (term frequency-inverse document frequency [25]). One thousand stopwords were used with no lemmatization. The relevance training corpus has 441 elements (3,512 words), split into two classes ("Relevant Document/Keep" and "Not relevant document/Discard").

The evaluation of the relevance was carried out separately and is calculated in terms of precision, recall and F-score. The evaluation presented in Table 2 was obtained using 10-fold cross validation; "Macro" represents the average of the values calculated for each class; "Micro" represents the global score, regardless of classes.

Table 2. Evaluation of textual relevance.

	Precision	Recall	F-score
Micro	0.925	0.925	0.925
Macro	0.928	0.931	0.921

The precision, recall and F-score are high (>0.90): the proposed method is able to detect the relevant documents correctly.

4.2 NERC

Evaluation was carried out using the MUC-5 specification on the annotated corpus described in paragraph 2. We provide here the evaluation result of the best model for each entity type. As it can be seen in Table 3, we can achieve the best results for all of the classes except one (*Network Type*) by combination of two models, the first of which has precedence in case of conflict.

Table 3. Results for each category of entities, using the best performing systems. [e] indicates the use of an ensemble model, where the first system has priority.

Entity	Occurrences	Precision	Recall	F-score	Model
Network_type	360	70.41	68.54	69.46	corenlp
Treatment	342	66.51	62.77	64.59	[e]corenlp-spaCy
Network_element	367	61.86	65.20	63.49	[e]spaCy-corenlp
Works	310	59.74	67.29	63.29	[e]spaCy-corenlp
Spatial	1001	58.05	67.75	62.52	[e]spaCy-corenlp
Measure	882	59.07	65.42	62.08	[e]corenlp-spaCy
Temporal	219	42.53	74.50	54.15	[e]heidel-corenlp
Water_quality	136	51.85	53.85	52.83	[e]corenlp-spaCy
Network_characteristics	196	55.00	47.06	50.72	[e]spaCy-corenlp
Network_management	255	40.47	42.83	41.61	[e]corenlp-spaCy
Indicator	592	35.90	33.26	34.53	[e]corenlp-spaCy
Organization	129	44.23	27.06	33.58	[e]corenlp-spaCy
Accident	26	34.78	20.00	25.40	[e]corenlp-spaCy

The current results are encouraging for a subset of categories, however they still require improvement for other ones, for which the trained models seem to perform not as well. In some cases, the lower results are probably caused by the fact that some categories, such as *Accident*, are poorly represented in our corpus. In other cases, such as for *Indicator*, the problem may be linked to heterogeneity of this category, which includes modifiers such as adverbs and adjectives related to various types of information (temporal, spatial, domain specific...).

4.3 Relation extraction

As we have seen, the Relation Extraction (RE) module is rule based and adds semantic links between the various entities in order to identify units of knowledge. In order to eliminate any noise caused by possible NERC errors, and thus to evaluate RE performances in isolation, a sub-set of the gold standard was automatically annotated with relations; the output contains 2,913 relations. Seven documents corresponding to a sampling rate of 30% were randomly selected for expert evaluation. At this stage only precision was evaluated, and experts checked automatically extracted relations between entities, assessing and labelling them as correct or incorrect. Missing relations were not taken into account. The rate of correctly detected relations is relatively high (precision = 0.83). Errors mostly occur when Named Entities, *i.e.* spatial ones, are juxtaposed. Linkage results are also impacted by errors in Named Entity recognition and text conversion, *i.e.* missing punctuation marks which modify the sentence structure and impact dependency rules.

5 Use case and Discussion

The WEIR-P information extraction pipeline offers users the possibility of rapidly acquiring information on the wastewater network of a city. The pipeline is a time-saver namely because the automatic Web-scraping phase, using a pre-set list of keywords, enables the user to multi-task, while the relevance check reduces the number of documents the user has to go over. In addition, the entity recognition module improves visual foraging and reduces reading time as a highlighted text will be more likely to be attended to and remembered than a plain text [8]. Note that it takes 5 hours to run 393 queries, under a minute to determine the relevance of a corpus containing 1,040 documents, and 5 hours and 17 minutes to extract 147,423 entities on 534 documents classified as relevant. In comparison, the average silent reading rate of an adult is 238 words per minute [4]. It would thus take an average adult 10 days, 8 hours and 9 minutes non stop to merely read the 1,040 documents.

The platform is an interesting aid as it can reconstruct events. Thus, hotspots can be identified through the "Accident" or "Works" labels or simply by analysing the frequency of occurrence of street or district names. A fully automated process is being implemented in the new version of the pipeline to carry out this task. Indeed, this type of information may also be useful to wastewater network managers who have just been granted concessions by public authorities in a new city and are not yet familiar with the network's history. Two representatives of the private sector leaders in Computer-aided Maintenance Management Systems (CMMS)/Enterprise Asset Management (EAM) and water and wastewater treatment services, are part of the project's steering committee and have been following the pipeline's development. Both expressed high interest in the pipeline's ability to recover dates and link them with network equipment as it would help plan maintenance operations. This feature was also highlighted by a representative of the public service in charge of wastewater management. We

performed a test on the city of Montpellier and were able to recover 233 occurrences of the word "pose" (to lay, to place in French), 559 of "mise en place" (implementation), 512 of "extension" and 375 occurrences of "réhabilitation" in the "Works" category. This type of information may also be recovered using digital archives uploaded manually by the user into the pipeline in pdf or txt format. The implementation of the platform would not be costly to local stakeholders as small material and human resources are needed to run it. The day to day life of the institutions and their regular business practices would not be affected as it would be mostly used for asset management i.e. for decision making at mid-management level. However, as with any new tool, training and time will be necessary to take in the change in practice.

As with many emerging tools, there is of course room for improvement. The evaluation results show that the current version of the system still presents various shortcomings that will be addressed in an improved version of the pipeline. In terms of information content, quantitative data (geometry, hydraulic performance) is mentioned less than events (i.e. works or accidents) in our documents (3,333 occurrences *vs.* 11,936 in the Montpellier Corpus). Also, the granularity of the spatial data is based on the type of document: street or district names are often mentioned in both technical reports and newspaper articles, however real-world coordinates are seldom found in the latter. Thus the WEIR-P pipeline may be a good tool to complete existing network databases and GIS systems. This would imply using data fusion techniques to combine and merge sometimes conflicting information. In order to improve on the current system, a larger manually annotated corpus may be necessary. The genre adaptation of NERC, exploiting the results of the text classification, will also be implemented.

A Sample-based generalization strategy [31] is implemented to ensure the genericity of the tool. Since WEIR-P relies heavily on Statistical learning, new samples from other French cities are currently being used for training. Validation is being carried out on other French speaking countries. The first tests on the city of Abidjan (Ivory Coast) are encouraging. The NER module is able to correctly label the network elements and the relation linking module will undergo further training in order to take into account local language uses.

6 Conclusions and perspectives

We have presented a global model of the information extraction from documents related to wastewater management and a platform which implements it. The preliminary results obtained on the Montpellier corpus are encouraging and show how a mix of supervised and rule-based techniques can be used to extract useful information and reconstruct the various phases of the extension of a given wastewater management network. The pipeline may also be used to recover the dates when given pieces of equipment were laid. This feature is deemed very useful by managers who need to plan ahead maintenance operations for old assets with missing implementation dates. The genericity of the tool we have developed is being assessed through tests on other cities in France and in French Speaking

countries. Indeed, some countries in North and West Africa, namely Morocco and Ivory Coast use Special Technical Specifications (STS) guidelines that are strongly inspired by the French ones. A quick analysis of the guidelines used in Quebec [21] shows that the technical vocabulary used to designate the network elements are also similar to the French ones. However, local language uses may vary. For instance the expression "assainissement d'un quartier" (sanitation of a neighbourhood) refers to sanitation/wastewater network laying for the French media and to cleaning of illegal occupation of public space for the Abidjanese. Thus, some adaptation work might be necessary to remove ambiguities.

The system still requires slight improvements as the information extraction pipeline produces some noise. Manual inspection by an expert of the extracted results may therefore be necessary and could be carried out using the visualization and spatial representation modules which enable users to easily assess the extracted data and further improve the models. In order to improve on the current system, a larger manually annotated corpus may be necessary. We also plan to use alternative classification algorithms such as One-class SVM that has been successfully adapted for reduced training samples [17] and perform semantic relation extraction based on document textual genre.

We believe that the domain modelling work carried out within MeDo will be useful to others working in the same domain, on French as well as on other languages. Since the NLP systems used in the NERC module of our pipeline support multiple languages, we assume that their adaptation should be a straightforward procedure.

References

1. Altaweel, M., Bone, C.: Applying content analysis for investigating the reporting of water issues. Computers, Environment and Urban Systems **36**(6), 599–613 (2012). https://doi.org/10.1016/j.compenvurbsys.2012.03.004
2. Ananiadou, S., Pyysalo, S., Tsujii, J., Kell, D.B.: Event extraction for systems biology by text mining the literature. Trends in Biotechnology **28**(7), 381–390 (2010). https://doi.org/10.1016/j.tibtech.2010.04.005
3. Araya, F., Faust, K., Kaminsky, J.A.: Understanding hosting communities as a stakeholder in the provision of water and wastewater services to displaced persons. Sustainable Cities and Society **57**(November 2019), 102114 (2020). https://doi.org/10.1016/j.scs.2020.102114
4. Brysbaert, M.: How many words do we read per minute? a review and meta-analysis of reading rate. Journal of Memory and Language **10**, 104047 (2019). https://doi.org/https://doi.org/10.1016/j.jml.2019.104047
5. Chahinian, N., Delenne, C., Commandre, B., Derras, M., Deruelle, L., Bailly, J.S.: Automatic mapping of urban wastewater networks based on manhole cover locations. Computers, Environment and Urban Systems **78**, 101370 (2019). https://doi.org/10.1016/j.compenvurbsys.2019.101370, https://hal.archives-ouvertes.fr/hal-02275903
6. Chahinian, N., Bonnabaud La Bruyère, T., Conrad, S., Delenne, C., Frontini, F., Panckhurst, R., Roche, M., Sautot, L., Deruelle, L., Teisseire, M.: Gold Standard du projet MeDo. https://doi.org/10.23708/H0VXH0, DataSuds, V1 (2020)

7. Chahinian, N., Bonnabaud La Bruyère, T., Delenne, C., Frontini, F., Panckhurst, R., Roche, M., Sautot, L., Teisseire, M.: Guide d'annotation du projet MeDo. https://doi.org/10.23708/DAAKF1, DataSuds, V1 (2020). https://doi.org/10.23708/DAAKF1
8. Chi, E.H., Gumbrecht, M., Hong, L.: Visual foraging of highlighted text: An eye-tracking study. In: Jacko, J.A. (ed.) Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments. pp. 589–598. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
9. Chinchor, N., Sundheim, B.: Muc-5 evaluation metrics. In: Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993. pp. 25–27 (1993)
10. Cookey, P.E., Darnsawasdi, R., Ratanachai, C.: Text mining analysis of institutional fit of Lake Basin water governance. Ecological Indicators **72**, 640–658 (2017). https://doi.org/10.1016/j.ecolind.2016.08.057
11. Copara, J., Knafou, J., Naderi, N., Moro, C., Ruch, P., Teodoro, D.: Contextualized French Language Models for Biomedical Named Entity Recognition. In: Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes. pp. 36–48. ATALA et AFCP, Nancy, France (2020)
12. COVADIS: Standard de données réseaux d'AEP & d'assainissement, version 1.2 (2019), http://www.geoinformations.developpement-durable.gouv.fr/
13. Dominguès, C., Jolivet, L., Brando, C., Cargill, M.: Place and Sentiment-based Life story Analysis. Revue française des sciences de l'information et de la communication **-**(17), 0–22 (2019). https://doi.org/10.4000/rfsic.7228
14. Ekstrom, J.A., Lau, G.T.: Exploratory text mining of ocean law to measure overlapping agency and jurisdictional authority. In: Proceedings of the 2008 International Conference on Digital Government Research. p. 53–62. dg.o '08, Digital Government Society of North America (2008)
15. Explosion: spacy. https://spacy.io/ (2019)
16. Gregory, I.N., Hardie, A.: Visual GISting: bringing together corpus linguistics and Geographical Information Systems. Literary and Linguistic Computing **26**(3), 297–314 (2011). https://doi.org/10.1093/llc/fqr022
17. Guerbai, Y., Chibani, Y., Hadjadji, B.: The effective use of the one-class svm classifier for handwritten signature verification based on writer-independent parameters. Pattern Recognition **48**(1), 103 – 113 (2015). https://doi.org/10.1016/j.patcog.2014.07.016
18. Hori, S.: An exploratory analysis of the text mining of news articles about water and society. In: Brebbia, C.A. (ed.) WIT Transactions on The Built Environment, vol. 1, pp. 501–508. WIT Press (2015). https://doi.org/10.2495/SD150441
19. Kergosien, E., Farvardin, A., Teisseire, M., Bessagnet, M.N., Schöpfel, J., Chaudiron, S., Jacquemin, B., Lacayrelle, A., Roche, M., Sallaberry, C., Tonneau, J.P.: Automatic Identification of Research Fields in Scientific Papers. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 1902–1907. European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12, 2018 2018)

20. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), http://www.aclweb.org/anthology/P/P14/P14-5010
21. Ministère du Développement Durable de l'Envionnement et de la lutte contre les changements climatiques: Description des ouvrages municipaux d'assainissement des eaux usées (DOMAEU) - Guide de rédaction. Tech. rep., Direction générale des politiques de l'eau, Direction des eaux usées (2018), https://www.environnement.gouv.qc.ca/eau/eaux-usees/ouvrages-municipaux/domaeu-guide-redaction.pdf
22. Park, K., Okudan-Kremer, G.: Text mining-based categorization and user perspective analysis of environmental sustainability indicators for manufacturing and service systems. Ecological Indicators **72**, 803–820 (2017). https://doi.org/10.1016/j.ecolind.2016.08.027
23. Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V., Bragge, J.: The design science research process: A model for producing and presenting information systems research. In: Proceedings of First International Conference on Design Science Research in Information Systems and Technology DESRIST (2006)
24. Rogers, C., Hao, T., Costello, S., Burrow, M., Metje, N., Chapman, D., Parker, J., Armitage, R., Anspach, J., Muggleton, J., Foo, K., Wang, P., Pennock, S., Atkins, P., Swingler, S., Cohn, A., Goddard, K., Lewin, P., Orlando, G., Redfern, M., Royal, A., Saul, A.: Condition assessment of the surface and buried infrastructure/-a proposal for integration. Tunnelling and Underground Space Technology **28**, 202 – 211 (2012). https://doi.org/10.1016/j.tust.2011.10.012
25. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., USA (1986)
26. Schmid, H.: Treetagger, a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart **43**, 28 (1995)
27. Shardlow, M., Nguyen, N., Owen, G., O'Donovan, C., Leach, A., McNaught, J., Turner, S., Ananiadou, S.: A new corpus to support text mining for the curation of metabolites in the ChEBI database. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 280–285. European Language Resources Association (ELRA), Miyazaki, Japan (may 2018), https://www.aclweb.org/anthology/L18-1042
28. Strötgen, J., Gertz, M.: Heideltime: High quality rule-based extraction and normalization of temporal expressions. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 321–324. Association for Computational Linguistics (2010)
29. Venkata Sailaja, N., Padmasree, L., Mangathayaru, N.: Incremental learning for text categorization using rough set boundary based optimized Support Vector Neural Network. Data Technologies and Applications **54**(5), 585–601 (2020). https://doi.org/10.1108/DTA-03-2020-0071
30. Wang, W., Stewart, K.: Spatiotemporal and semantic information extraction from Web news reports about natural hazards. Computers, Environment and Urban Systems **50**, 30–40 (2015). https://doi.org/10.1016/j.compenvurbsys.2014.11.001
31. Wieringa, R., Daneva, M.: Six strategies for generalizing software engineering theories. Science of Computer Programming **101**, 136–152 (2015). https://doi.org/10.1016/j.scico.2014.11.013, http://dx.doi.org/10.1016/j.scico.2014.11.013